

# Developing Automated Amplified Fragment Length Polymorphism Analysis to Identify Microbial Species and Strains

Paul J. Jackson  
Los Alamos National Laboratory  
505-667-2775  
jackson@telomere.lanl.gov

Co-Investigators:  
Karen K. Hill, Richard T. Okinaka and Lawrence O. Ticknor  
Los Alamos National Laboratory  
Paul Keim  
Northern Arizona University

## Objective

The primary objective of this project is to develop an automated, rapid, user-friendly approach to identify pathogenic microbes to the species and strain level. Amplified Fragment Length Polymorphism (AFLP) analysis provides an excellent method of rapidly interrogating a microbial genome to provide phylogenetic information. The method permits analysis of many more genetic loci than is possible using other methods, providing significantly more resolution than methods that rely on comparative DNA sequencing of individual loci. Such an analysis allows discrimination below the species level. When genomic sequences are not available, it can also be used to identify loci that are the basis of DNA fingerprinting systems using multiple locus VNTR analysis (MLVA). To accomplish our goals, we must:

- Develop the necessary software to accurately read the AFLP profiles
- Tie this analysis to computational methods that interpret these phylogenetic data
- Populate an archive with AFLP profiles for all of the threat agents and phylogenetically related species

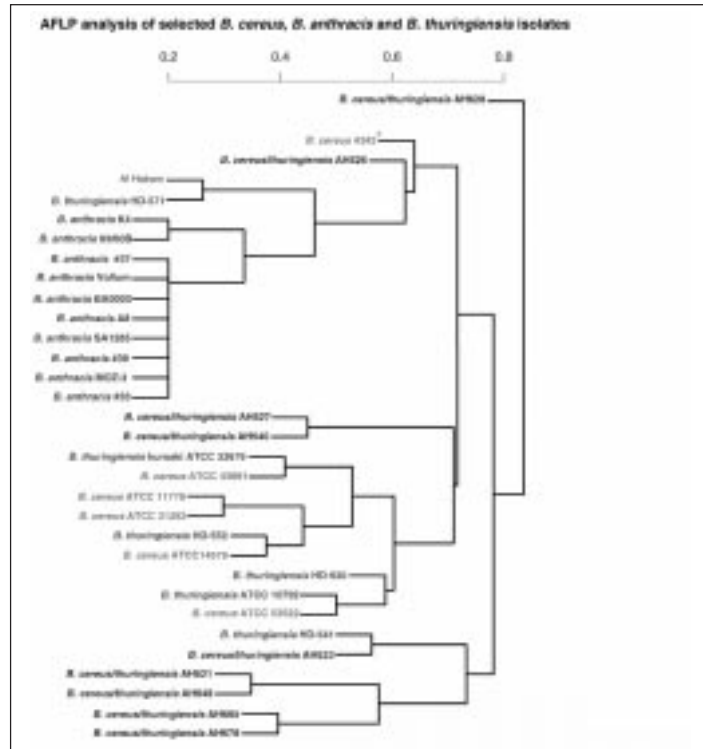
## Recent Progress

### Populating the Archive

AFLP analysis is limited by the number of available profiles from different microbial species and strains that are available for comparison to an uncharacterized sample and by the availability of computational methods that allow rapid comparison of a large population of different profiles and phylogenetic analysis based on these comparisons. We are developing AFLP profiles for specific threat agents and their closest phylogenetic relatives with an emphasis on those phylogenetically related microbes that most likely would interfere with a rapid identification of a biothreat agent.

We have almost completed AFLP analysis of a sufficient number of *Bacillus* species and strains to allow rapid placement of previously uncharacterized *Bacillus* isolates within a detailed phylogenetic tree of this genus. There is a clear distinction between any *B. anthracis* isolate and any other *Bacillus* characterized so far. However, our most recent detailed analyses demonstrate that *B. anthracis* is not genetically peculiar among the *B. cereus* subgroup but is closely related to some members of this group (see figure on following page). We have also completed AFLP analysis of all available *Yersinia pestis* isolates and conducted a comparison of these profiles to those of other closely related

This technique provides a clear distinction between any *B. anthracis* isolate and any other *Bacillus* characterized so far.



species. A similar analysis has been initiated on *Clostridium botulinum* and is only limited by the availability of representative strains of this species. There is a relationship between the phenotypic characters used to distinguish strains of this species and the AFLP profiles obtained for these. However, this relationship does not extend to the toxins present within an isolate. The presence of a specific toxin cannot necessarily be predicted by the phylogenetic information, suggesting that the ability to produce a specific toxin can be transferred laterally to other strains without transfer of a large portion of other genetic information.

## Computational Analysis

As the number of AFLP profiles to be compared increases, it is impossible to do this rapidly using manual methods. We have developed computation methods to compare large numbers of profiles and determine their phylogenetic relationships. Computational approaches are focusing on methods of rapidly comparing profiles to one another and to profiles generated for previously characterized microbial isolates. This is a difficult problem because of minor differences among different gel profiles and the inability of current software packages to accurately identify DNA fragment sizes. However, significant progress has been

made. It is now possible to rapidly compare profiles generated from a single new isolate to the large number of profiles in our archive and identify a small number of archived profiles that are similar or identical to the new profile. If an AFLP profile from a test sample that is represented in the database is compared to the archived profiles, 73% of the time the software identifies a single AFLP profile that matches, 16% of the time it identifies two possible matches and 4% of the time it eliminates all but three profiles.

### Technology Application

We have applied this technology to analysis of multiple microbial samples provided by other sponsors and collaborators. A thorough analysis of a large *Bacillus* collection using this archive and software provides new information about the relationship of *B. anthracis* to its closest relatives and about the phylogenetic relationships among all members of this *Bacillus* subgroup. The figure on the previous page demonstrates the utility of this approach to understand the complex relationships among different microbial species.

### Future Outlook

More work is required to continue populating the archive, to make the software more user-friendly, to improve the connection between the archive and the newly generated profiles, to improve resolution of the analysis algorithm, and to allow automated phylogenetic analysis based on the profile comparisons. However, we have now demonstrated that it is possible to automate the entire process. As computational methods have improved, it is evident that a small number of the profiles in our current archive are not satisfactory and must be reentered. We have developed a method that detects unsatisfactory AFLP signatures and will replace these profiles with newly generated data. We will also continue populating the archive with AFLP profiles for representative samples of all the other microbial threat agents. We have recently demonstrated that a modification of our current methods allows detection and characterization of viruses. When possible, we will generate and archive profiles for the different viral agents. When the viral genome sequence is known, we can predict the AFLP pattern. Deviations from this pattern suggest the isolation of new strains or intentional manipulation of the pathogen.

More work is needed on computational aspects of the project. Database search routines will be further polished and experimental AFLP data can be compared to “theoretical” AFLP results based on genomic sequences of some microbes. We can also improve the analysis routines including clustering, distance measures and tree techniques. We will also increase the search speed. We must also collect the available information on all the samples (i.e., source, geographic origin, pathogenic or

A thorough analysis of a large *Bacillus* collection using this archive and software provides new information about the relationship of *B. anthracis* to its closest relatives and about the phylogenetic relationships among all members of this *Bacillus* subgroup.

other phenotypic characteristics) in the archive and develop the computational methods to rapidly tie this to the phylogenetic analysis. As more samples become available from other sources, these will be analyzed and compared to the current archive to better understand these samples and to better understand the phylogenetic relationships among pathogenic and nonpathogenic microbes. All newly analyzed samples will themselves become part of the archive so that collection of specific strains from different sources can rapidly be detected.

AFLP analysis provides information about which DNA fragments are most variable among different closely related species and among different strains of the same species. This information can be used to identify DNA fragments that most probably will be species-specific. The DNA sequences of such fragments will then provide information for development of pathogen-specific PCR primers. Fragments that vary among different strains of the same species provide a source of information for development of PCR primers that distinguish among different strains of the same species. AFLP analysis requires analysis of purified DNA from a single source. PCR analysis with strain-specific primers provides information of complex sample content without the necessity of purifying a single microbe from the complex mixture. Such primers have been used to analyze complex forensic and environmental samples.

## Publications

- P. Keim, A. Klevytska, L.B. Price, J.M. Schupp, G. Zinser, R. Okinaka, K.K. Hill, P.J. Jackson, K.L. Smith, M.E. Hugh-Jones, "Molecular Diversity in *Bacillus anthracis*," *Journal of Applied Microbiology* 87, 215–217 (1999).
- P.J. Jackson, K.K. Hill, M.T. Laker, L.O. Ticknor, P. Keim, "Genetic Comparison of *B. anthracis* and its Close Relatives Using Amplified Fragment Length Polymorphism and Polymerase Chain Reaction Analysis," *Journal of Applied Microbiology* 87, 263–269 (1999).
- R.T. Okinaka, K. Cloud, O. Hampton, A.R. Hoffmaster, K.K. Hill, P. Keim, T.M. Koehler, G. Lamke, S. Kumano, J. Mahillon, D. Manter, Y. Martinez, D. Ricke, R. Svensson, P.J. Jackson, "Sequence and Organization of pX01, the Large *Bacillus anthracis* Plasmid Harboring the Anthrax Toxin Genes," *Journal of Bacteriology* 181, 6509–6515 (1999).
- R. Okinaka, K. Cloud, O. Hampton, A. Hoffmaster, K. Hill P. Keim, T. Koehler, G. Lamke, S. Kumano, D. Manter Y. Martinez, D. Ricke, R. Svensson, P.J. Jackson, "Sequence, Assembly and Analysis of pX01 and pX02," *Journal of Applied Microbiology* 87, 261–262 (1999).

## Development of Genetic Signatures for Identification and Typing of Biological Threat Agents

Gary Andersen  
Lawrence Livermore National Laboratory  
925-423-2525  
andersen2@llnl.gov

### Objectives

Rapid and accurate detection and typing of biological threat agents are essential to respond appropriately to the release of potentially harmful microorganisms. These approaches are important not only for more immediate measures, but also for later forensic analysis and attribution. The purpose of this program is to find and exploit differences in the genetic material among BW agents and their relatives to allow rapid and sensitive detection and typing. These differences can be used to accurately discriminate among similar microorganisms, a crucial aspect of an accurate test. Two major types of discrimination are desirable. The first is the ability to distinguish threat agents from very similar harmless microbes that are often normally found in the environment. Erroneous detection of benign close relatives without the ability to clearly distinguish them from a biological weapon would lead to false results thus rendering the test useless. At the same time, the test must allow detection of the threat agent even though there can be considerable diversity within species of microorganisms. In other words, a test that misses detection of a given agent on occasion is of limited value. The second type of discrimination is the ability to group microorganisms within a species once an organism has been identified. This ability is essential for forensic analysis and attribution. Even though they may cause the same disease, strains within a species can vary based on such factors as geographical origin and the time taken from the environment (e.g., an infected host) into the laboratory. A test that can exploit these differences can produce patterns or signatures that can be used to match or distinguish different isolates of a particular agent. This information could be used to narrow the likely sources of a particular strain. Also, a released agent could be matched with a high level of confidence to organisms acquired through subsequent legal investigation.

Using the genetic material (nucleic acids) of microorganisms as a basis for signatures has several advantages over other typing methods. Most often, the genetic material is in the form of DNA, but in the case of some viruses is in the form of the less stable DNA relative, RNA. This material is often very stable and is present even when the organisms are no longer alive. Also, nucleic acid detection methods such as the polymerase chain reaction (PCR) are extremely sensitive, relatively easy to optimize compared to other methods, highly reproducible, and the test-specific materials necessary are easy to generate and store. The principal goals of the project are:

- Identification of genetic differences among microbial threat agents and non-pathogenic relatives
- Identification of genetic differences among different groups within a particular agent

Using a technique known as subtractive hybridization, DNA from a given microorganism can be compared to DNA from a close relative and segments that are unique to the pathogenic agent in question can be isolated

- Database analysis to define signatures based on these differences
- Development of laboratory protocols to facilitate the use of such signatures
- Development of high-throughput methods to increase efficiency, better methods of database analysis, and a central database for the standardized use of multiple laboratories and organizations.

## Recent Progress

Using a technique known as subtractive hybridization, DNA from a given microorganism can be compared to DNA from a close relative and segments that are unique to the pathogenic agent in question can be isolated by the removal of segments that are common. In the case of some viruses that use RNA as the genetic material, the RNA can be converted to DNA for the purposes of these experiments. The specific structure, or sequence, of the unique DNA segments can be determined and based on this information another technique, the polymerase chain reaction (PCR), can be used to detect the presence of minute amounts of a given segment and therefore the organism which carries it. In any such comparison, these segments may or may not correlate well with a given agent but rather may be found in another close relative or may be absent in other sources of the pathogen. Therefore, the presence or absence of the various DNA segments must be tested in a wide variety of strains and closely related species to find those segments that are truly agent-specific. Similarly, strains within a species can be grouped into different subspecies or biovars based on a variety of biological tests. When strains are compared by subtractive hybridization, PCR is used to test a wide variety of strains and closely related species to find biovar specific markers. The results of all these tests are then combined and analyzed to find candidate signatures. Subtractive hybridization has many advantages over other approaches to signature development. Small amounts of DNA are adequate and are often available from collaborators, thus obviating the need for specific expertise of a given organism. The genetic makeup of the organisms does not need to be characterized. Finally, the likelihood of developing specific signatures is maximized as all the genetic material of the organism participates in the comparison, rather than other methods that focus on specific regions.

This process has been applied to compare two closely related pathogens, *Yersinia enterocolitica* and *Yersinia pseudotuberculosis*, both close relatives of the plague agent, *Yersinia pestis*. Although these bacteria are highly similar, they cause different diseases, are found in different hosts, and are transmitted differently. In spite of the similarities, multiple subtractive hybridization experiments have yielded a great

many markers specific to either *Y. enterocolitica* or *Y. pseudotuberculosis*. In addition, strain specific markers were also identified within each of these species. These studies more recently have been expanded to include *Yersinia pestis*. Various strains of *Y. pestis* originating from different locations around the world were compared and strain-specific signatures were developed. Many of the markers were also tested against *Y. enterocolitica* and *Y. pseudotuberculosis*, yielding pestis-specific markers. As a powerful and dramatic demonstration of the validity of the overall method, these experiments led to a single test that could both detect and distinguish these three species without detecting closely related species such as *Y. pestoides*.

There are a great many subspecies within *Salmonella enterica*, including the pathogenic strains that are well-known causes of food-borne illnesses. One such subspecies, enteritidis, is now the most common cause of salmonellosis in the U.S., and was used to deliberately infect 751 people in Oregon in the 1980's in an attempt to influence voter turnout in a local election. Individuals are normally infected through undercooked chicken eggs. The above procedures have been used to find enteritidis-specific diagnostic markers that identify all but a few rare strains of enteritidis and do not detect an extensive array of the many closely related biovars that are common in the environment but do not infect eggs. In addition, these markers are not detected in the great many close relatives of the genus *Salmonella* that are found in the same environments. In addition to the potential benefits of providing an excellent diagnostic test for monitoring egg production, these results again demonstrate the power of these techniques because in spite of the large number of closely related species and subspecies, a well-tailored signature was developed.

Concomitant with the achievements described above, the efficiency of these approaches has been greatly improved through automation and increased parallel processing of samples. This will spur an increased rate of growth of the signature database.

## Future Outlook

Further progress in signature development can be divided into three main areas:

- Additional increases in efficiency of the techniques.
- Development of a central database as well as improved methods for computer data analysis.
- Continued development of signatures for particular threat organisms.



## Homology between *Bacillus anthracis* Plasmid-Encoded Genes and Other Bacterial Species

Cheryl R. Kuske  
Los Alamos National Laboratory  
505-665-4800  
kuske@lanl.gov

Co-Investigators:  
John M. Dunbar, James A. Pannucci and Richard Okinaka  
Los Alamos National Laboratory

### Objectives

The goal of this project is to identify regions of gene homology between genes encoded on the pX01 and pX02 plasmids of *B. anthracis*, other *Bacillus spp.* common in the environment, and other pathogenic bacteria, to improve specificity of DNA-based detection methods for *B. anthracis* in environmental samples, and to determine potential gene function of new genes. We are using a combined DNA hybridization and PCR approach to identify regions of homology between the known *B. anthracis* virulence genes and closely related non-pathogenic *Bacillus spp.* to identify regions of the genes that are unique to the pathogen. Recent sequencing of the pX01 and pX02 plasmids (funded by this program) has identified over 200 potential genes (open reading frames) for which no function can be assigned. Using DNA hybridization and PCR assays, we are identifying those open reading frames that are conserved between *B. anthracis*, closely related *Bacillus spp.*, and other bacterial pathogens.

### Recent Progress

#### Known Virulence Genes

Our first objective was to identify bacterial genes and specific regions of those genes that had homology at the DNA or amino acid level to any of five well-characterized *B. anthracis* vir genes (pag, lef, cya, capA, capB, capC). PCR primers and hybridization probes that included conserved regions of sequence were designed using homology information from database searches. Several primer pairs were designed for each target gene and tested extensively in PCR reactions for ability to amplify the target gene and potential homologs.

Dot blot and Southern blot hybridization experiments were conducted to identify gene homology between the *B. anthracis* vir genes and over 40 non-pathogenic *Bacillus* species that are potentially common in different environments. Homology between the known *B. anthracis* vir genes and multiple species of nonpathogenic *Bacilli* was detected at low stringency hybridizations. We sequenced the 16S rDNA gene of several of the positive species and conducted phylogenetic analysis to identify whether the hybridizing species were closely related to *B. anthracis*. Species hybridizing with the vir genes were widely dispersed among the *Bacilli*, indicating that these genes may be present in very divergent species. Hybridizing sequences from the other species are being cloned and sequenced to identify the extent of homology.

#### Environmental Testing

DNA from a wide variety of soil and other environmental samples was surveyed for the presence of *B. anthracis* vir genes using PCR. The vir gene homology study was initiated because we had observed considerable cross reaction between some *B. anthracis* specific PCR primers and a couple of environmental DNAs that we knew did not contain the pathogen. We have completed a survey of over 20 environmental samples using PCR

Genes present on *B. anthracis* virulence plasmids are also present in nonpathogenic *Bacillus* species that are common in the environment. Investigations are under way to identify the extent of shared genes between the pathogen and nonpathogenic relatives. This work will help in design of DNA-based detection strategies that are specific for the pathogen.

amplification of primers for the *cap* and *cya* genes. In this survey, we have only detected false positives between the *B. anthracis* genes and the native microflora in one sample, which is encouraging.

#### New pX01 and pX02 Open Reading Frames

We are using a combined hybridization and PCR approach to determine whether new genes on pX01 and pX02 for which no function can be assigned are specific to *B. anthracis* or are shared between the pathogen, other *Bacilli*, and other pathogenic bacterial species. PCR primers have been generated for about half of the 143 open reading frames on pX01, and each of these genes has been amplified to generate 1-kb length probes that are being used to screen DNA from a panel of 12 *Bacillus* spp. The hybridization assays have identified homology primarily in *B. cereus* 43881 and *B. thuringiensis* 33679 (kurstaki). PCR amplification and gene sequencing the DNA from the hybridizing species has confirmed the hybridization results and may help assign possible functions to some of these new genes. For example, gene fragments amplified from *B. cereus* 43881 or *B. thuringiensis* 33679 show 89–95% similarity to pX01. One region of similarity brackets a possible origin of replication for pX01. Experiments to determine the genomic location (plasmid or chromosome) of the homologous genes in these two species are in progress. Analysis of pX01 is complete, and we are now conducting a similar analysis of pX02 genes.

### Future Outlook

We expect to complete this project in FY00. Milestones include:

- Finishing the cloning and sequence analysis of homologs to known virulence genes
- Completing comparative analysis of new pX01 and pX02 genes that have no assigned function
- Preparation and submission of manuscripts describing this work.

# Development of Pathogen DNA Fingerprinting Systems Using Multiple Locus Variable Number Tandem Repeat Analysis

Paul Keim  
Northern Arizona University  
520-523-1078  
paul.keim@nau.edu

## Objective

High-resolution molecular typing of pathogens can provide (1) geographical correlation with strain types, (2) dissection of epidemics, and (3) precise genetic “matches” for forensic attribution. In order to discriminate among strains of different bacterial pathogens, we are developing multiple variable number tandem repeat (VNTR) markers. These are the most variable regions in bacterial genomes and offer the greatest discriminatory power for DNA fingerprinting. *Bacillus anthracis* is the most homogeneous bacteria described, yet VNTR markers detect great diversity among strains. When multiple VNTR loci are combined in an analysis, the potential molecular typing power is increased exponentially. The goal of this work is to develop automated and highly discriminatory markers systems in *B. anthracis* and *Yersinia pestis* for forensic and attribution applications.

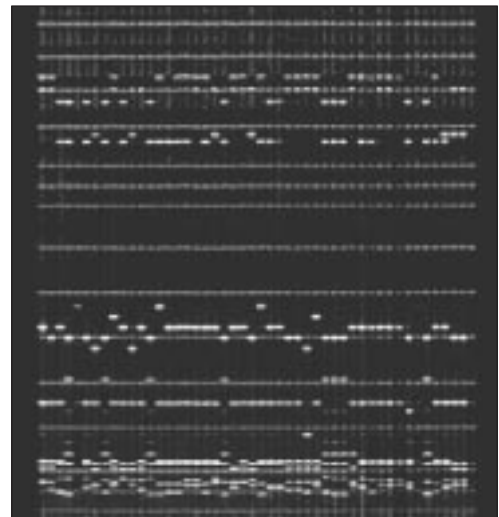
## Recent Progress

We have developed multiple locus VNTR analysis (MLVA) systems for *B. anthracis* and *Y. pestis* and used these systems to analyze worldwide diversity patterns. These MLVA systems use multiplexed PCR reactions and are automated through the use of fluorescent labels and standard genotyping software from Applied Biosystems. The data are standardized for easy transfer between laboratories and for database development.

### Anthrax

The MLVA system for *B. anthracis* contains eight VNTR marker loci. These markers are a mixture of simple sequence repeats, complex repeats, and chromosomal and plasmid loci. These markers were developed from variable AFLP fragments or from plasmid nucleotide sequences. Fluorescent PCR primers of three different colors allow for markers of the same size to be simultaneously analyzed (see the figure on the right). In addition, amplicons vary in size such that all eight markers loci are analyzed in a single electrophoretic lane. Various combination of PCR primers have lead us to an optimum of three multiplex reactions to minimize amplifications. Automated genotyping is facilitated by “macros” that use the fluorescent color and amplicon size to identify the marker loci and alleles.

The anthrax data are being maintained in a custom “Access” database that facilitates analyses and provides a ready dissemination instrument. The Microsoft Access 97 database software



We have developed multiple locus VNTR analysis (MLVA) systems for *B. anthracis* and *Y. pestis* and used these systems to analyze worldwide diversity patterns.

is readily available to any laboratory with a personal computer. The Anthrax Genotypic Database consists of 25 items: 18 forms, 3 tables, and 4 query windows. The 18 forms also have numerous queries as integrated functions. There are 44 fields, including information about specific isolates ranging from collection site/source to genotype data to the precise location of each sample in the Keim laboratory. No prior database experience is necessary to use this database. The database is expandable to handle future samples and MLVA markers.

New VNTR markers are being discovered and tested for suitability for inclusion in the next-generation MLVA diagnostic system. Six additional VNTR loci were identified from AFLP markers and are being tested for diversity. With the development of the *B. anthracis* genome sequencing project by (TIGR), we have developed bioinformatic approaches for identifying potential VNTRs from genomic sequence. Over 800 potentially variable genomic structures have been identified, and 60 have been chosen to test for variability.

Over 400 unique isolates have been analyzed from all parts of the world, except the former Soviet Union. The world collection can be subdivided into 89 unique genotypes with the eight marker MLVA system. The genotypes cluster into approximately six major groups, probably representing clonal lineage. Particular genotypes are found to dominate major anthrax outbreak's, though the actual diversity pattern is a function of each outbreak's history. We have performed molecular epidemiological studies in North America, South Africa and Australia.

### Plague

The MLVA system for *Yersinia pestis* contained 19 VNTR marker loci, as of September 1999. All of the markers are simple sequence repeats located on the chromosome (as opposed to plasmid-borne markers). The diversity of individual markers is greater, on average, than those in the *B. anthracis* system possibly due to the simple nature of the repeats. As with the *B. anthracis* system, we have developed four multiplex reactions (7, 5, 4 & 3 loci per mix) to detect variation at the 19 loci. "Macros" have been written for automated genotyping. A plague database is currently under construction, modeled after the anthrax database. We have identified over 800 additional potential VNTR loci and are screening these for utility in next-generation MLVA system.

A small set of strains representing world wide diversity has been analyzed, as well as 95 samples collected across California over the past 20 years. These genotypes are being analyzed for spatial and temporal variation patterns. We can detect the affect of geographical and temporal distance on sample relationships.

## Future Outlook

Our new bioinformatic methods for identifying informative VNTRs have proven extremely effective and has eliminated one of the most problematic aspects of MLVA development: VNTR discovery. This major advancement was achieved during the FY99 period and foresees great progress in the future. Over the next grant period we will be using the novel VNTRs to fine tune the MLVA systems in *B. anthracis* and *Y. pestis*, while developing novel MLVA systems in other pathogens including *Francisella tularensis*.

### Technology Transfer to Other Federal Agencies

In order to facilitate the transfer of the MLVA systems, we are conducting short training courses for personnel from other agencies. These consist of a one-week regime where technically skilled individuals learn wet bench procedures, instrument handling, data analysis and database interfacing. The first training course was conducted for the Centers for Disease Control and Prevention (CDC). In collaboration with our laboratory, the CDC will use the MLVA system to DNA fingerprint their entire strain collection (>1000 isolates).

## Publications

- D.M. Adair, P.L. Worsham, K.K. Hill, A.M. Klevytska, P.J. Jackson, A.M. Friedlander, P. Keim. "Diversity in a Variable Numbers of Tandem Repeat (VNTR) from *Yersinia pestis*," *Journal of Clinical Microbiology* (2000, In press).
- P.J. Jackson, K.K. Hill, M.T. Laker, L.O. Ticknor, P. Keim, "Genetic Comparison of *B. anthracis* and its Close Relatives Using AFLP and PCR Analysis," *Journal of Applied Microbiology* 87, 263–269 (1999).
- P. Keim, A. Klevytska, L.B. Price, J.M. Schupp, G. Zinser, R. Okinaka, K. Hill, P.J. Jackson, K. Smith, M. Hugh-Jones, "Molecular Diversity in *Bacillus anthracis*," *Journal of Applied Microbiology* 87, 215–217 (1999).
- P. Keim, L.B. Price, A.M. Klevytska, K.L. Smith, J.M. Schupp, R. Okinaka, P.J. Jackson, M.E. Hugh-Jones, "Multiple-Locus VNTR Analysis (MLVA) Reveals Genetic Relationships within *Bacillus anthracis*," in review with *Journal of Bacteriology* (2000).
- R. Okinaka, K. Cloud, O. Hampton, A. Hoffmaster, K. Hill, P. Keim, T. Koehler, G. Lamke, S. Kumano, D. Manter, Y. Martinez, D. Ricke, R. Svensson, P.J. Jackson, "Sequence, Assembly and Analysis of pX01 and pX02," *Journal of Applied Microbiology* 87, 261–262 (1999).

- R. Okinaka, K. Cloud, O. Hampton, A. Hoffmaster, K. Hill, P. Keim, T. Koehler, G. Lamke, S. Kumano, J. Mahillon, D. Manter, Y. Martinez, D. Ricke, R. Svensson, P. Jackson, "The Structure and Organization of pXO1, the Toxin Containing Plasmid of *Bacillus anthracis*," *Journal of Bacteriology* 181, 6509–6515 (1999).
- J.M Schupp, L.B. Price, A. Klevytska, P. Keim, "Internal and Flanking Sequence from AFLP Fragments Using Ligation-Mediated Suppression PCR," *Biotechniques*, 26, 905–910 (1999).
- K.L. Smith, V. DeVos, H. Bryden, M.E. Hugh-Jones, P. Keim, L.B. Price, A. Klevytska, D.T. Scholl, "Meso-scale Ecology of Anthrax in Southern Africa: a Pilot Study of Diversity and Clustering," *Journal of Applied Microbiology* 87 (1999).

## Multiple Locus Sequence Typing: A Phylogenetic Approach for Defining Signatures

Richard Okinaka  
Los Alamos National Laboratory  
505-667-2743  
okinaka@telomere.lanl.gov

Co-investigators:  
Rita Svensson and Paul J. Jackson  
Los Alamos National Laboratory

Paul Keim  
Northern Arizona University

### Objective

This study utilizes comparative DNA sequencing strategies to differentiate bacterial pathogens. Multi-locus sequence typing (MLST) combines the polymerase chain reaction (PCR) and DNA sequencing to gather precise sequence information from multiple regions of the genomes of bacteria and other organisms. This information can be used to distinguish closely related species and individuals at the species and subspecies level. The method targets the analysis of common regions in related species and is therefore similar to ribosomal gene (rDNA) analysis that has been used to establish the phylogenetic relationships among organisms residing in the microbial world. MLST expands these analyses to multiple regions that evolve more quickly than rDNA to build larger datasets that greatly increases genetic resolution, particularly for close relatives that have little or no variation in their ribosomal DNA sequence. MLST sequence information from pathogens and their close relatives can now be electronically stored, transferred and shared between multiple laboratories. This technology provides a sequenced based method for comparing broad regions of the genomes of related species and provides reagents for future DNA microarray and minisequencing approaches to type and identify strains and species of pathogens (e.g., see J. Nolan, Detection Technologies).

### Recent Progress

The MLST approach has been used to analyze *Bacillus anthracis* and some of its close relatives, *Bacillus thuringiensis*, *Bacillus cereus* and *Bacillus mycoides*. PCR primers have been designed and tested to amplify approximately 500-bp fragments from 15 distinct regions from all of these genomes. Each of these regions were then amplified and sequenced from at least 25 strains of *B. anthracis*, 15 strains of *B. cereus*/*B. thuringiensis* and one strain of *B. mycoides*. These analyses consistently indicate that the *B. anthracis* sequences are extremely homogeneous and that clear distinctions can be made among each of the *B. anthracis* amplicons and those of *B. cereus*/*B. thuringiensis*. For example, eight positions containing single nucleotide polymorphisms (SNPs) have been identified in a 466-bp region of the *rpoC* gene. All *B. anthracis* strains (25) are identical while one or more of these sites differ among every strain of *B. cereus*/*B. thuringiensis*/*B. mycoides* analyzed. Two general conclusions can be made from both AFLP and MLST analysis of these species: *B. anthracis* appears to have evolved from *B. cereus* and *B. thuringiensis* and there are certain strains of *B. cereus* and *B. thuringiensis* that are very closely aligned with the *B. anthracis* sequences. These SNPs combined with those from 10 other genomic regions provide a large and powerful data set that

A direct DNA sequencing approach has been used to compare 15 regions of the *Bacillus anthracis* genome to those of its closest relatives.

These studies provide reagents and a database that can be used in the forensic analysis of a specific class of *Bacilli*.

can be used to unequivocally distinguish *B. anthracis* from its closest relatives in a phylogenetically meaningful fashion.

We have begun to extend these analysis to three other pathogens: *Burkholderia mallei*, *Clostridium botulinum* and *Coxiella Burnetii*. Small insert libraries have been constructed from ATCC strains of *Burkholderia mallei* and *Clostridium botulinum* and clones derived from these libraries are currently being sequenced. These sequences are being utilized to generate PCR primers that can be used for DNA amplification and sequencing of target sites for MLST analysis.

## Future Outlook

Single nucleotide polymorphic markers can be readily coupled to high-throughput identification assays including DNA microarrays, oligoligation assays, and other mini-sequencing methodologies. MLST analysis is a powerful tool for identifying molecular differences in closely related species and subspecies. The data generated is most useful for establishing the phylogenetic relationship among species that are difficult to distinguish by rRNA gene analysis. Current studies are aimed at developing MLST markers for at least three additional potential BW pathogens.

## Publications

- P. Keim, A. Klevytska, L.B. Price, J.M. Schupp, G. Zinser, R. Okinaka, K.K. Hill, P. Jackson, K.L. Smith, M.E. Hugh-Jones, "Molecular diversity in *Bacillus anthracis*," *Journal of Applied Microbiology* 87, 215–217 (1999).
- R. T. Okinaka, K. Cloud, O. Hampton, A.R. Hoffmaster, K.K. Hill, P. Keim, T.M. Koehler, G. Lamke, S. Kumano, J. Mahillon, D. Manter, Y. Martinez, D. Ricke, R. Svennson, P. J. Jackson, "The Sequence and Organization of pX01, the Large *Bacillus anthracis* Plasmid Harboring the Anthrax Toxin Genes," *Journal of Bacteriology* 181, 6509–6515 (1999).

Genbank entry accession numbers: AF065404, AF188935.

## Development of a Technique Based on Insertion Sequences for the Strain Identification of *Yersinia pestis*

Emilio Garcia  
Lawrence Livermore National Laboratory  
925-422-8002  
garcia12@llnl.gov

Vladimir Motin  
Lawrence Livermore National Laboratory  
925-424-4347  
motin1@llnl.gov

### Objective

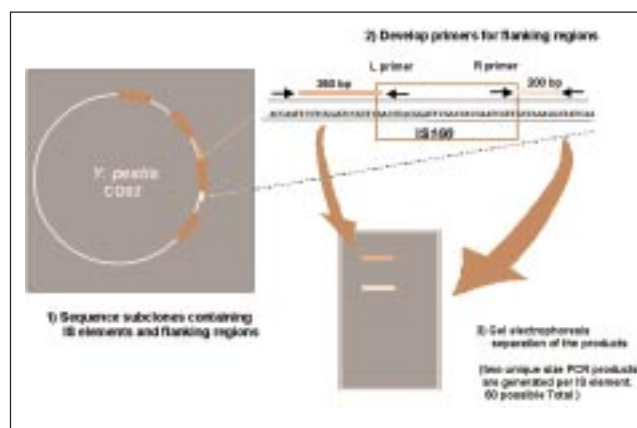
We are developing a genome-wide fingerprinting method for *Y. pestis* based on the chromosomal localization of a unique insertion element, IS100. This fingerprinting method should permit unambiguous identification of *Y. pestis* at the species level and, most important, it should enable the differentiation of geographically distinct strains of this organism.

The fingerprinting method makes use of our knowledge of the DNA sequence that surround each and everyone of the IS100 elements in the genome and the ability to amplify by PCR fragments that are unique to the IS100-neighboring gene boundary. The fingerprints obtained are highly informative with respect to the genomic organization of each strain, are rapidly obtained, and are fully automatable. The resulting set of unique fragments associated with a given strain provides the fingerprint or molecular signature for that strain (see the figure below).

To maximize the information obtained by this fingerprinting we have designed two primer pairs that localize the IS100 elements in the pMT1 plasmid, one pair that localizes the single element found in pPCP and the remaining primers localize the chromosomal copies of IS100. Thus, this type of test not only makes it possible to determine the genomic variation among the tested strains but also the plasmid composition.

Data have been accumulated that indicates that a unique molecular fingerprint based on the observed diversity of IS100 distribution among *Y. pestis* enables the differentiation of strains representing almost every large geographical location in the world including all biovars of *Y. pestis*.

The fingerprinting data obtained up to now indicate that it is possible to distinguish strains according their biovar classification, plasmid composition and geographical origin. As one would expect, strains isolated from the United States give a fingerprint that is associated with the biovar Orientalis (plague is believed to have arrived to the United States from China or somewhere in Asia). Similarly, KIM, a strain originally isolated in Kurdistan and Iran and belonging to the biovar Medievalis, gives a fingerprint characteristic to this biovar. The IS-fingerprinting is amenable to multiplexing and it does not require availability of pure cultures (it works on mixed samples). Although the work performed up to



This technique makes use of intrinsic variability to obtain quickly and inexpensively a “fingerprint” signature that both identifies and helps to geographically locate the origin of a given *Y. pestis* isolate.

now has only included the use of 15 primer pairs to generate a fingerprint pattern, in the subsequent years we will expand the primer pairs utilized as well as the number of stains that will be fingerprinted. Our final goal is to accumulate a fingerprint database that contains a very large worldwide collection of *Y. pestis* strains and primers capable of localizing all possible positions in which the IS100 element can insert itself (regardless of strain).

The aim of this research is to develop a fingerprint methodology that will enable the unambiguous identification of *Yersinia pestis* strains and their geographical attribution. The development of DNA-based identification and detection tools for *Y. pestis* poses some important challenges. The presence of near-neighbors closely related at the genetic level, the sharing of an important virulence plasmid with related strains, and the fact that *Y. pestis* strains isolated from different geographical origin tend to contain atypical plasmids hamper the development of simple methods for its molecular identification. In addition, *Y. pestis* strains contain a large number of insertion sequence elements (ISs) scattered throughout their genomes which are highly variable with respect to their location. The methodology being developed makes use of this intrinsic variability to obtain quickly and inexpensively a “fingerprint” signature that both identifies and helps to geographically locate the origin of a given *Y. pestis* isolate. Because the method is PCR-based and ready automatable it should be readily transferable for use on specialized detectors being developed under the “Detection” section of this program located on page 65.

## Recent Progress

Using a set of 17 PCR primer pairs that localize a number of IS100 elements on the chromosome of *Y. pestis* strain CO92, we have developed a fingerprinting method that allows rapid species and stain identification of members of this group.

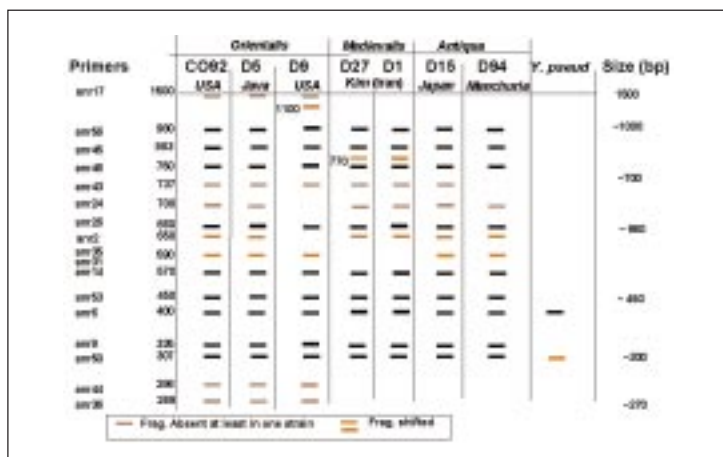
We have accumulated data that indicate that a unique molecular fingerprint is obtained based on the observed diversity of IS100 distribution among *Y. pestis* strains and other pathogenic *Yersinia* species.

The fingerprinting technique has been applied to a collection of some 112 distinct *Y. pestis* strains corresponding to:

- Different biovars (Orientalis, Medievalis and Antiqua)
- Distinct geographical origin (representing a worldwide distribution)
- A set having undergone multiple laboratory manipulations and passages

Typical fingerprints obtained for members of each biovar are depicted on the figure below. Multiple laboratory passages in vitro and in vivo did not change the IS100 fingerprint profile in any of the analyzed strains.

We have obtained consensus fingerprint associated with the biovar *Orientalis*.



Finally, in a finding that may constitute the most important development of this year, we have obtained preliminary evidence that a Manchurian strain belonging to the Antiqua biovar displays a typical pattern of the *Orientalis* biovar. Such a finding suggests that such “Antiqua” isolate could represent an ancestor of the *Y. pestis* clone responsible for the third (and present) plague pandemic.

## Future Outlook

In the coming year we expect to get a consensus fingerprint for the biovars Antiqua and Medievalis. Such consensus will enable the unequivocal placement of *Y. pestis* strains to a given biovar and possibly to a geographical location.

We will automate our present fingerprinting methodology by using an automated fluorescent-detecting DNA sequencer (ABI377) available at our facility. Such instruments, in conjunction with its accompanying genotyping software (GenScan™ and Genotyper™) enables rapid capture of large amounts of data which can be easily manipulated and standardized. The production of standardized data from this readily available type of instrument will enable cross-laboratory comparisons and will facilitate the construction of generalized databases for pathogen signatures.

We will adapt the IS100-based fingerprinting technology for use with detection technologies being developed by the “Detection” effort of the CBNP (i.e., fluorescent micro-spheres, TacMan™, high-density arrays, etc.).

We will construct genomic libraries of each biovar representative that will enable us to localize and then catalogue each possible position on the chromosome in which the IS100 element can incorporate. With that capability we could fingerprint and cross-reference each an every *Y. pestis* isolate regardless of its origin

We will try to expand a collaboration with the Russian Anti-Plague Institute in Saratov, Russia, which will enable the fingerprinting of their extensive *Y. pestis* collection. This year we have initiated with them the sequencing of a *Y. pestis*-specific phage that may enable us to identify the gene(s) responsible for host-specificity and viral attachment to the *Y. pestis* cell. Such research could guide our effort to construct in the future non-antibody-based detectors with exquisite specificity (this phage has been shown to attach and lyse thousands of *Y. pestis* isolates from the Russian collection at Saratov but never a *Y. pseudotuberculosis* strain).

We plan to obtain additional information to corroborate our preliminary finding regarding the possible ancestor to the Orientalis biovar. Such finding would do much to elucidate the mechanisms of virulence evolution. This work will complement nicely with the whole genome comparative sequencing of *Y. pseudotuberculosis*, that aims to understand the evolution of *Y. pestis* (an obligate pathogen) from its immediate ancestor, *Y. pseudotuberculosis* a free-living facultative pathogen of man.

## Expression Studies of Virulence Factors in *Yersinia pestis*

Emilio Garcia  
Lawrence Livermore National Laboratory  
925-422-8002  
garcia12@llnl.gov

Vladimir Motin  
Lawrence Livermore National Laboratory  
925-424-4347  
motin1@llnl.gov

### Objective

The aim of this research is to identify a battery of new genes in *Yersinia pestis* that participate in the virulence process. The methodology being developed makes use of newly acquired DNA sequence from this organism to construct a high-density array or DNA chip that will enable massive and global analyses of the expression pattern of the entire genome of *Y. pestis*. Because the majority of the technology required for this project is available either commercially or through collaborators and our group and collaborators, have substantial knowledge on the biology of *Y. pestis*, the project is very likely to yield a substantial amount of new data on the virulence process in this organism.

Discovery of new virulence factors in *Y. pestis* will directly impact the development of new signatures for detection and attribution of this organism. It will also enable us to utilize the new technology to study other pathogens such as *Brucella*, for which there is substantially less knowledge of the virulence process. Our preliminary activities in the area of expression chip development has already led us to make some important contacts with scientist working on the biology of *Brucella* and *Francisella*, two organisms on which we plan to begin studies in the coming years.

### Recent Progress

#### Isolation and Characterization of KatY

Using purified protein isolated from *Y. pestis* cells subject to temperature shift, we have characterized a new thermoregulated, chromosomally-encoded catalase-peroxidase (KatY) which is also present in *Y. pseudotuberculosis* but not in *Y. enterocolitica*. The approach employed anti-KatY monoclonal antibodies to screen a *Y. pestis* genomic library to isolate a 13 Kb insert of a positive clone. The entire insert was of sequenced and the putative peptides encoded by the open reading frames were compared with the N-terminal sequence of KatY determined by protein sequencing.

The predicted KatY gene consisted of 737 amino acid residues, possessed a prokaryotic signal sequence of 23 amino acids, and contained the motif typical of other bacterial catalase-peroxidases. Interestingly, the promoter region of the KatY contained three repeats that showed a significant homology with the consensus sequence recognized by LcrF, the transcription activator of the *Yersinia* virulence regulon. This finding is consistent with the previously described observation that KatY is synthesized by *Yersinia* during expression of the low-calcium response. The time course of the expression of KatY was determined by dot-blot hybridization with the corresponding

100%

100%

the first task for this project next year will be optimization of conditions for deposition and detection on the proposed DNA chip. One of the major technical roadblocks for expression monitoring in bacteria is the difficulty encountered in isolating sufficient quantities of mRNA from these organisms. We will be adapting to *Y. pestis* mRNA isolation techniques developed for *E. coli* by collaborators at the University of California at Berkeley.

The actual first expression studies on *Y. pestis* will involve the monitoring of some 200 genes encoded in all three virulence plasmids of this organism. Such studies are of great importance because they will test our ability to monitor the expression of 50 or more known virulence genes in this organism. This work will constitute the pillar on which our subsequent work will be based since it will guide us in identifying potential new virulence genes on the basis of their co-regulation. This work will also enable us to determine the temporal expression of individual components of the type III secretion apparatus in this organism as well as all genes of unknown function encoded on the plasmids. Elucidation of the temporal functioning of this very important and ubiquitous virulence system will provide a significant contribution to the understanding of the present virulence model of *Yersinia* and several other pathogens that share this pathway (i.e., *Pseudomonas*, *Salmonella*, *Shigella*, etc).

After the work described above, we plan to perform global gene expression experiments that will include the approximately 4,500 genes encoded in the chromosome of *Y. pestis*. These experiments will be carried out under various physiological conditions including those encountered in organisms recovered from infected human cell lines. This type of experiment has not been possible until now and, together with experiments carried out directly from *Y. pestis* cells recovered from the organs of infected animals, are one of the most important experiments that will become possible from the development of this new technology. We expect to conduct the latter part of this work (FY01 and FY02) in collaboration with Russian collaborators from the Anti-Plague Institute in Saratov, Russia and/or with our collaborator at the Pasteur Institute in Paris, France.

Towards the second year of this project we plan to begin exploratory work on a similar expression chip for the BW agent *Brucella*. We have recently established an informal collaboration with investigators from Fullerton State University and a group in Argentina working on *Brucella abortus*. They have provided us with the DNA sequence for some 2,000 genes of this organism that will enable us to place them on a chip and carry out similar expression analyses. This work will allow us to begin work on a new BW pathogen while incorporating technology developed during the our studies of *Yersinia pestis*.

## Publications

E. Garcia, Y.A. Nedialkov, J.M. Elliott, V.L. Motin, R. Brubaker, "Molecular Characterization of Cloned KatY (antigen 5), a Thermoregulated Chromosomally Encoded Catalase-peroxidase of *Yersinia pestis*," *Journal of Bacteriology* 181, 3114–3122 (1999).



## Toxin and Virulence Factor Structure/Function Determination and Protein Signature Development

Rod Balhorn  
Lawrence Livermore National Laboratory  
925-422-6284  
balhorn2@llnl.gov

Subramanyam Swaminathan  
Brookhaven National Laboratory  
516-344-3187  
swami@bnl.gov

Diana Roe  
Sandia National Laboratories  
925-294-4905  
dcroe@ca.sandia.gov

Co-Investigators:  
S. Kadkhodayan, M. Knapp, F. Lightstone,  
B. Rupp, B. Segelke and S. Ringhofer  
Lawrence Livermore National Laboratory

M. Young and H. Adalsteinsson  
Sandia National Laboratories

S. Eswaramoorthy and Y. B. Zhang  
Brookhaven National Laboratory

### Objective

This program provides the high resolution structural information on key protein toxins, virulence factors, and protein determinants unique to organisms or spores that are needed to enable the development of rapid, sensitive methods for the detection and identification of threat biological agents using protein-based signatures. The molecular structures of these proteins and the complexes they form with important receptors are identified using a combination of X-ray diffraction, Nuclear Magnetic Resonance spectroscopy and mass spectrometry. This information is being used to direct the design and synthesis of robust, highly specific molecular reagents that can be used in place of antibodies to detect toxins and organisms in environmental samples and exposed individuals with high fidelity and sensitivity. The primary goals of this effort are to

- Identify protein toxins and proteins produced by threat organisms that can be used to detect agent dispersal or confirm human exposures
- Crystallize and determine the molecular structure of key toxin proteins, virulence factors, and organism specific proteins
- Design suite of robust “2<sup>nd</sup> Generation” molecular reagents that bind to toxins and virulence factors with high affinity and selectivity
- Develop analogous organism specific reagents to replace antibodies and to complement DNA based detection methods

Future efforts will use combinatorial methods to identify molecules that bind specifically to the surface of pathogenic bacteria and their spores for use in organism detection.

### Recent Progress

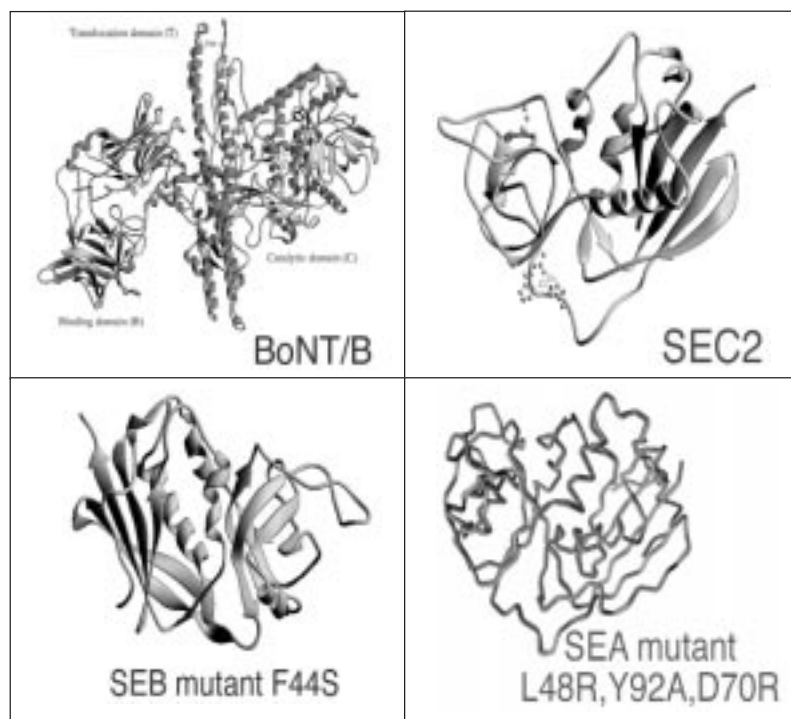
Working with collaborators at USAMRIID, USAMRICD, NIH, VA Medical Center (Pittsburgh) and Walter Reed Army Hospital, we have crystallized and determined the structures of several toxins or toxin functional

The information provided by these structures is being used to elucidate how the toxin binds to cell surface receptors and identify critical sites on the toxin surface that lead to toxin:receptor recognition.

domains produced by Clostridial and Staphylococcal bacteria and complexes these proteins form with receptor molecules. Using computational docking techniques and mass spectrometry, we have also begun to identify a number of first generation molecules that bind to specific sites on the targeting (or receptor binding) domain of tetanus toxin and the Staphylococcal enterotoxin C3.

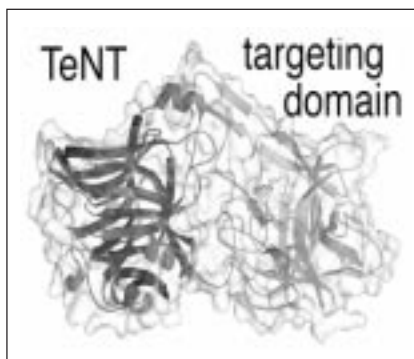
#### Determination of Toxin/Virulence Factor Structures

Five new toxin crystal structures were determined by X-ray diffraction: the tetanus toxin (TeNT) targeting domain (a homolog of the BoNT domains), intact botulinum neurotoxin B (BoNT/B), *Staphylococcus aureus* enterotoxin C2 (SEC2), and two *S. aureus* enterotoxin vaccine



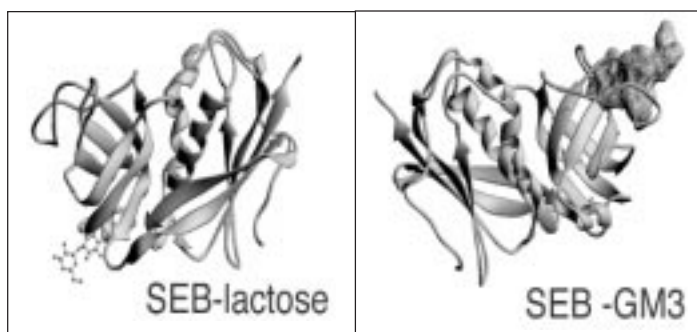
candidates, SEB mutant F44S and the SEA mutant L48R/Y92A/D70R. Three additional proteins, the catalytic domain of BoNT/A, intact BoNT/E, and NAP Hn-33, have been crystallized and current efforts are being focused on completing their structures. A virulence factor produced by *Yersinia pestis*, the V antigen, has been produced and has been screened to identify optimal crystallization conditions.

The high-resolution structure of the TeNT targeting domain (1.6Å) has been used to initiate the identification of a group of small molecules (“First Generation” reagents) that bind to specific sites on the surface of the protein using a computational technique called docking. This structure has also been used to predict the structures of a series of BoNT targeting domains by Comparative Modeling. These models, and the recently determined structures of the intact BoNT/A and BoNT/B toxins, are being used to identify conserved sites on BoNT for targeting the design of Clostridial neurotoxin-specific reagents.



#### Mode of Interaction with Receptors

Two crystal structures of the Staphylococcal enterotoxin SEB complexed with the carbohydrate portion of the ganglioside GM3 and a related sugar (lactose) were also completed. The information provided by these structures is being used to elucidate how the toxin binds to



cell surface receptors and identify critical sites on the toxin surface that lead to toxin:receptor recognition. Two additional complexes, BoNT/B with a small molecule inhibitor designed by USAMRICD and BoNT/B with the VAMP peptide, were also crystallized.

### Computational Modeling of Toxins, Ligand Identification by Docking, and Experimental Verification of Binding by Mass Spectrometry

Computational molecular recognition (CMR) tools are being developed to accelerate the design of new, robust molecular reagents that can replace antibodies in affinity-based toxin and organism sensors. A large database of small molecules (~240,000) is being screened by computational docking to identify new molecules that may bind to specific sites on the surfaces of protein toxins, virulence factors or other organism-specific protein determinants (such as receptors). Mass spectrometry is used to determine which of these molecules actually bind to the protein and to obtain an estimate of their binding affinity.

During the past year, a series of suitable “pockets” have been identified on the surfaces of the tetanus toxin targeting domain and the Staphylococcal enterotoxin C3 (SEC3), and sets of small molecules have been identified as possible “First Generation” ligands that might bind to these sites. Ligand sets were identified for binding to three sites on Tetanus toxin and one site on SEC3. Approximately half of the ligands have been screened for binding to two sites on the tetanus toxin targeting domain by electrospray mass spectrometry, and more than 50% were found to bind. The ligands projected to bind to Site 1, a possible ganglioside binding site, were also tested for their ability to compete with ganglioside for binding to the targeting domain using a ganglioside-liposome binding assay, and one of the seven ligands was found to compete for GT1b binding.

### Development of Computational Recognition Tools

To speed up the docking step, we have designed a computational framework, DEMoS (Distributed Extensible Molecular Simulator), which will be used to perform computational docking on clusters of distributed computers. This code currently has capabilities for Molecular Dynamics simulations of proteins and ligands, and the docking tools are being incorporated. The scoring scheme used in docking is being optimized by comparing the results obtained for a group of approximately 80 protein/ligand complexes with known binding affinities. Quantum chemical methods are also being developed to treat the electronic interactions of the ligand with the surrounding solvent.

### Design and Synthesis of “Second Generation” Reagents for the Detection and Identification of Threat Toxins and Organisms

By synthetically combining the best of each group of “First Generation” ligands that are identified to bind to neighboring sites on the surfaces of these proteins, we will create a new class or “Second Generation” of molecular reagents that can be used in place of antibodies for detecting toxins and other proteins with high affinity and selectivity. These

reagents will be provided to other groups to enable the development of more stable, robust detection systems for threat agents or protein signatures in exposed individuals with a substantially lower frequency of false positive signals. Our first synthetic schemes involve conjugating individual ligands to the ends of appropriate length spacer molecules to generate one bidentate molecule at a time.

## Future Outlook

By the end of FY00, we will have completed high-resolution structures of the free BoNT/A light chain, and several key complexes of BoNT with molecules that identify toxin recognition mechanisms and regions of the molecule required for receptor/substrate specificity. Work will be nearing completion on the solution of the complete structure of the final BoNT serotype, BoNT/E. Future work will focus on obtaining high resolution structures for the other serotype targeting domains to complete the Clostridial neurotoxin structures needed to identify conserved structural sites and to design the BoNT-specific Second Generation reagents. The structure of the V antigen of *Y. pestis* and edema factor of *B. anthracis* will be completed by early FY01, and efforts will turn to identifying and characterizing the structures of key bacterial and spore protein determinants that will aid the detection effort.

We will implement the boundary element method (BEM) for solvation into the Massively Parallel Quantum Chemistry code (MPQC), and algorithms using flexible docking will be compared with the regular docking procedure to determine if they provide a better prediction of the most stable binding sites in our docking studies. Scoring schemes used in the docking algorithms will also continue to be evaluated and improved using sets of known protein-ligand interactions.

The next computer modeling and docking studies will focus on the BoNT targeting and catalytic domains, using the coordinates provided by the structural studies to identify structural “pockets” and sets of ligands that might bind to those sites that appear to be conserved and essential (e.g., they cannot be altered by genetic engineering). Following the synthesis of the prototype bidentate reagent for TeNT detection to confirm the utility of the approach, an efficient combinatorial synthesis scheme will be developed to generate suites of these Second Generation reagents for detecting the BoNTs. This approach will then be extended to develop similar reagents for the Staphylococcal enterotoxins, ricin, protein components of anthrax toxin (for use in detecting/confirming anthrax infection), and other protein determinants produced by threat organisms.

## Publications

- S. Kadkhodayan, M.S. Knapp, J.J. Schmidt, S.E. Fabes, B. Rupp, R. Balhorn, "Cloning, Expression and One-Step Purification of the Minimal Essential Domain of the Light Chain of Botulinum Neurotoxin Type A," *Protein Expression & Purification* (in press 1999).
- F. C. Lightstone, M.C. Prieto, A.K. Singh, M.C. Piqueras, R.M. Whittall, M.S. Knapp, R. Balhorn, D.C. Roe, "The Identification of Novel Small Molecule Ligands that Bind to Tetanus Toxin," *Chemical Research in Toxicology* (submitted 1999).
- S. Swaminathan, W. Furey, M. Sax, "Structures of Staphylococcal Enterotoxin B complexed with Glycosphingolipid Saccharide Moieties," *Protein Science* (submitted 1999).
- S. Swaminathan and S. Eswaramoorthy, "Crystallization and Preliminary Studies on *Clostridium Botulinum* Neurotoxin Type B," *Acta Crystallographica Section D* (submitted 1999).
- S. Swaminathan and S. Eswaramoorthy, "Crystal Structure of *Clostridium Botulinum* Neurotoxin Type B at 1.8 Angstrom Resolution," *Structure* (submitted 1999).

## Signature Pattern Development for Detection of “Out-of-Place” Organisms

Catherine A. Macken  
Los Alamos National Laboratory  
505-665-6464  
cam@t10.lanl.gov

Tom Burr  
Los Alamos National Laboratory  
505-665-7865  
tburr@lanl.gov

Co-investigators:  
H. Lu, H. Mukundan, G. Myers and A. Skourikhine  
Los Alamos National Laboratory

### Objectives

BW agents possess distinct molecular features that can be used to monitor their natural occurrence in the environment. In order to distinguish a natural instance of an organism from a subtly modified, unnatural variant with enhanced virulence, we need (1) a molecular “signature” to characterize the natural organism’s background variation; and (2) a sensitive statistical test to recognize variants that differ significantly from this background.

In this project, we have focused on approaches to identify highly informative molecular signatures, using DNA and amino acid sequence data from viral organisms. Our recent goals were:

- Compare the statistical characteristics of three methods for defining signature patterns.
- Examine the sensitivity of the methods to sample sizes.

### Recent Progress

After our focus on hepatitis C earlier in this project, we moved to influenza as a model organism. Influenza has many characteristics of a potential BW agent: it is highly contagious and is spread as an aerosol; it can be very lethal, with death rates, such as during the 1918 pandemic, that far exceed any seen due to other diseases in the 20th century; a vaccine can be produced for protection of a rogue nation against a modified strain, but it takes too long to produce vaccine to protect a victimized nation; a few subtle changes, possibly as few as one amino acid change, can lead to dramatic changes in its virulence; several hemorrhagic fevers of BW concern, such as Lassa, Junin and Machupo, possess similar genomic structures and are also spread as aerosols; and epidemics caused by influenza bear the characteristics of rapid, wide spread that one might expect from release of an aerosol, contagious BW agent.

The genes of influenza mutate at a rapid rate. This variability makes it feasible, and efficient, to develop methods for detection of “out-of-place” influenza viruses on the basis of genomic sequence data.

Influenza is highly contagious and spread as an aerosol; it can be very lethal, with death rates that far exceed any seen due to other diseases in the 20th century.

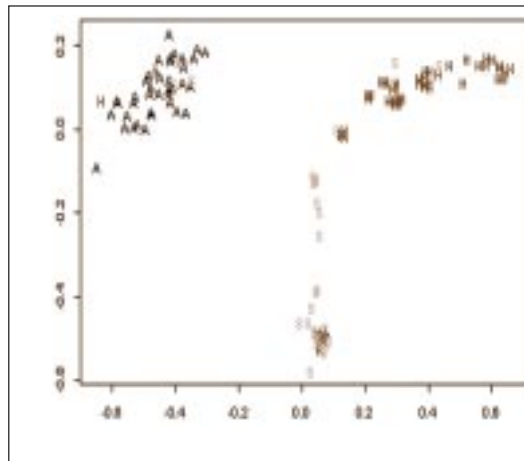
## Techniques for Pattern Recognition

We have followed two basic directions toward signature pattern development: supervised learning, and unsupervised learning. Supervised learning uses data from known classifications (the “training set”) to identify characteristics in the data that both strongly indicate class membership, and accurately distinguish among classes. We considered two methods of supervised learning: VESPA and principal components analysis (PCA). The latter method is standard in the literature for quantitative characteristics; our contribution was to extend its use to categorical data. VESPA was developed in earlier years’ efforts under this program

Unsupervised learning assumes no prior knowledge of classifications. A distinguishing pattern must be determined by examining data in different ways, in search of a view that clearly delineates meaningful groupings. The unsupervised technique that we focused on was phylogenetic tree reconstruction, which depends upon a model of the evolutionary behavior of the influenza virus.

### Supervised Learning

Principal components analysis (PCA) attempts to define two or a small number of orthogonal axes that capture as much of the variability in the sample as possible, such that the predefined classes overlap as little as possible. In the figure below, sequences from avian, human and swine hosts are plotted on the axes given by the first two principal components. To generate the figure below, genetic distances among sequences were estimated under a simple model of evolution. Samples were then plotted on (x,y) coordinates such that Euclidean distances among the sequences were as close as possible to the corresponding



genetic distances, thus achieving a huge reduction in the dimensionality of the problem.

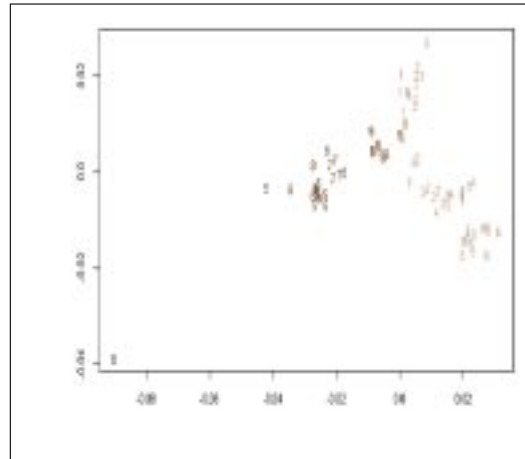
These two axes allow a clean clustering of sequences by host, with 14 cases out of 129 being misclassified. Upon researching the sources of the misclassified cases, we learned that all were instances of cross-species transmission, and that the

host of origin belonged to the class with which it was clustered. For example, the “human” sequence in the “avian” class was taken from a person infected from poultry.

By exploiting the derived functional relationship between genetic distance and (x,y) coordinates, we can place a novel sequence on this map, and assign it to a class with a high degree of certainty.

The figure below illustrates PCA applied to human influenza sequences from a 30-year time period. Our goal here is to determine a genetic signature of “place” in time. While it is difficult with this amount of data to distinguish years, it is clear that sequences can be clustered (somewhat crudely) by decades.

VESPA is not illustrated here, since this work was described in earlier reports. When applied to the dataset for PCA above, the classification results were exactly the same.



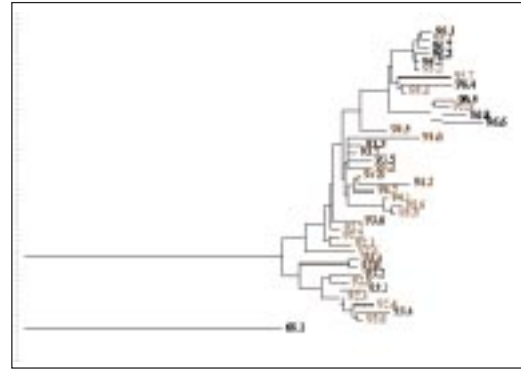
Unsupervised  
Learning

The techniques  
described above  
share attributes of

simplicity, and they share the detraction of falsely assuming that all sequences are independent. Thus, these techniques are susceptible to vagaries of a possibly nonrepresentative sample. Phylogenetic tree estimation is an important fundamental technique, designed to capture relationships by evolutionary descent among nucleotide or amino acid sequences. We used phylogenetic trees to classify sequences by “clade” or distinct lineage. Sequences that are members of a clade are more closely related genetically to each other than to sequences from other clades.

To estimate a phylogenetic tree, one needs a model of the evolution of the organism. We explored a hierarchy of complexity of models of evolution, using a sample of 80 human influenza viral sequences from the years 1968–1995. While details of the branch lengths of the tree were sensitive to the model, clustering of sequences changed little between models of moderate complexity and models of maximum complexity. Because computational time increases dramatically with complexity, we

based our development on models of evolution with moderate complexity, namely the F84 model with site-specific rates of evolution. We applied this mode of evolution to sequences from 1992–1996, with an additional sequence from 1968, the date at which the viral lineage purportedly arose. The results are given in the figure below. The sequences from 1992–1996 are color-coded by year. It is clear that significant overlap of lineages exist, indicating the need for more data in order to improve the power of discrimination.



### Which Technique?

For the datasets considered here, all techniques had a similar power to correctly assign samples to classes. However, we were operating under fortunate circumstances of ready access to large datasets, so that we could cull sequences and achieve even representation over time and within classes. We also could predefine a training set. In practice, these two predisposing circumstances may not be in place. It is our opinion that approaches to classification or detection of “out-of-place” events should be based on phylogenetic tree estimation, to reduce sensitivity to nonrepresentative data sets. The disadvantage of these techniques is their computational intensity, a disadvantage that is rapidly diminishing as supercomputers become readily accessible.

## Publications

- T. Burr and C.A. Macken, “Molecular Epidemiology of Influenza: A Model Biological Warfare Agent.,” LA-UR-99-3679 (1999).
- T. Burr, C. Macken, W. Bruno, A. Skourikhine, “Confidence Measures for Evolutionary Trees: Applications to Molecular Epidemiology.,” *IEEE Intelligence in Neural and Biological Systems* (1999, in press).

## Cooperative Epidemiology: The Nexus of Biological Weapons Proliferation and Emerging Diseases

Alan Zelicoff  
Sandia National Laboratories  
505-844-8020  
apzelic@sandia.gov

Serge Netesov  
State Center for Virologic Research (VECTOR), Russia

Gennady Lepyoshkin  
National Center for Biotechnology  
of Kazhakstan (NCB-RK)

Co-investigators:  
Arthurine Breckenridge  
Sandia National Laboratories

Gary Simpson  
New Mexico Department of Health

### Objectives

Under the CBNP in FY98-99, a one year pilot-scale regional monitoring network that incorporates advanced telecommunications technology was carried out in Chelyabinsk, Russia in cooperation with hospitals in New Mexico to ascertain the prevalence of and risk factors for acquisition of Hepatitis C. In this follow-on work, we are investigating the same disease in Russia (near Novosibirsk) and in northern Kazhakstan (near Stepnogorsk). The project will exercise epidemiological tools, information collation and analysis, and information distribution, which could be of great value in monitoring for the Biological Weapons Convention (BWC). We seek to increase the ability of regional public health departments in the former Soviet Union (FSU) to provide clinical expertise in identifying and reporting the rapidly changing patterns of this disease.

### Background

The spread of emerging diseases constitutes an international public health emergency. Understanding the origin, clinical effects, and prevalence of these novel or recurring diseases is essential for planning intervention strategies. At the same time, unusual disease outbreaks may represent evidence of biological weapons proliferation; investigation of such outbreaks could increase confidence that such episodes arise from natural events, as well as develop the cooperative means for addressing concerns of biological weapons proliferation around the world.

As mandated by the President, under current BWC negotiations, the U.S. will seek to strengthen existing confidence building measures (CBMs) and to negotiate on-site inspections under two specific conditions:

- Suspicious outbreaks of disease.
- Allegations of biological weapons use.

Further, the U.S. has sought to open Russian biological weapons laboratories to international scrutiny. The current phase of the Hepatitis C project will take place with the State Center for Virologic Research in Novosibirsk,

The morbidity and mortality from “emerging diseases” is enormous - Hepatitis C alone affects over 3 million Americans, and untold millions more around the world.

The Cooperative Epidemiology project has demonstrated that complex, novel diseases can be investigated by non-specialist physicians and provide information of importance to health care providers, while establishing a network of data gathering that could be exploited to resolve allegations of biological weapons use or development.

Russia (also known as VECTOR) and the National Center for Biotechnology of Kazakhstan (NCB-RK) in Stepnogorsk, Kazakhstan.

In order to address the President's mandates and to provide a near-term tool to meet the needs of the legally binding regime under negotiation, a networking tool for exchange of information is desperately needed. Current CBMs are filed by less than 20% of signatories to the BWC, and those that are filed are in paper format, which are largely unusable by States Parties.

Emergency room and clinic populations will be utilized for this project, as they provide populations which are largely unselected, a problem common to previous Hepatitis C surveys that have focused on known risk groups. Our experience from Chelyabinsk-70 suggests that there are novel risk groups, although the data are still being analyzed. This information has been factored into the risk-factor questionnaire being developed for this portion of the project.

Approximately 2000 volunteer patients over the age of 18 from the Russian and Kazhak sites will have serological studies for Hepatitis C performed. They will also respond to a detailed questionnaire of known and possible risk factors for Hepatitis C, designed with the assistance of the World Health Organization and the University of New Mexico School of Medicine. All data will be entered by participating centers. Serology will be performed using standardized test kits at each of the sites. In addition, laboratories will maintain serum samples for possible later analysis of genomic structure of Hepatitis C isolates. All data will be transmitted over the Internet to a central server at Sandia National Laboratories and will be accessible to all parties at all times.

The final phase of this work will be to publish the results in an internationally recognized epidemiology journal. Data will be shared with the international negotiating community in Geneva. All data will be authenticated and patient confidentiality protected. Those patients with positive results will be provided counseling for mitigation of Hepatitis C disease progression.

## Recent Progress

Data gathering was completed from approximately 2000 patients in the Chelyabinsk region and from roughly 1200 emergency room patients in three hospitals in New Mexico. This study is almost certainly the largest of its kind ever performed, providing a large amount of data on known and potential risk factors for acquisition of Hepatitis C.

Although data analysis is still preliminary, we have learned the following:

- The prevalence of Hepatitis C is much higher in the central Urals region of Russia than expected.
- It appears that ingestion of potential hepato-toxins (certain medications and also alcohol) is associated with Hepatitis C.
- Surgery, but not necessarily blood transfusions, is associated with Hepatitis C acquisition in Russia.

Russian and Kazhak biological weapons laboratories in both the civilian and military sector have submitted research proposals similar to this work. We believe that this is an important milestone in the conversion of these institutes to legitimate purposes.

## Next steps

We anticipate developing similar surveys for real-time outbreak monitoring of acute respiratory illness, beginning with influenza and evolving to include significant pediatric infectious disease as well as novel hemorrhagic fever syndromes.

In addition, we will:

- Develop project plans for real-time monitoring of influenza epidemics in northeastern Asia. It is highly likely that new strains (or variants) of influenza arise each year in this region of the world. This is critical information for development of annual vaccination strategies.
- Negotiate with the Kirov biological weapons laboratory for similar work. If we are successful, this will be the first time that a military bio-weapons site will engage in collaborative research with U.S. laboratories.
- Seek additional funding from the Centers for Disease Control and Prevention for permanent laboratory and clinical disease surveys at multiple FSU biological and public health laboratories.